Vendittoa I, Luis AS, Rydahl M, Schückel J, Fernandes VO, Vidal-Melgosa S, Bule P, Goyal A, Pires VMR, Dourado CG, Ferreira LMA, Coutinho PM, Henrissat B, Knox JP, Baslé A, Najmudin S, Gilbert HJ, Willats WGT, Fontes CMGA. Complexity of the *Ruminococcus flavefaciens* cellulosome reflects an expansion in glycan recognition. *Proceedings of the National Academy of Sciences of the United States of America* 2016, 113(26), 7136–7141.

The complexity of the *Ruminococcus flavefaciens* cellulosome reflects an expansion in glycan recognition

Immacolata Venditto[a,b*], Ana S. Luis[a,b,*], Maja G. Rydahl[c,*], Julia Schückel[c,*], Vânia O. Fernandes[a,d], Silvia Vidal Melgosa[c], Pedro Bule[a], Arun Goyal[e], Virginia M.R. Pires[a], Catarina G. Dourado[a], Luís M.A. Ferreira[a,d], Pedro M. Coutinho[f], Bernard Henrissat[f,g,h], J. Paul Knox[i], Arnaud Baslé[b], Shabir Najmudin[a], Harry J. Gilbert[b,1], William G. Willats[c,1] and Carlos M.G.A Fontes[a,d,1]

Running title: Novel CBM families in ruminal cellulosomes

Classification: BIOLOGICAL SCIENCES; Biochemistry

[a]*CIISA – Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal*; [b]*Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK;* [c]*Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Copenhagen, Denmark;* [d]*NZYTech genes & enzymes, 1649-038 Lisboa, Portugal;* [e]*Department of Biotechnology, Indian Institute of Technology Guwahati, Guwahati, Assam, India;* [f]*Architecture et Fonction des Macromolécules Biologiques, UMR7857 CNRS, Aix-Marseille University, F-13288 Marseille, France;* [g]*INRA, USC1408 Architecture et Fonction des Macromolécules Biologiques, F-13288 Marseille, France,* [h]*Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia;* [i]*Centre for Plant Sciences, University of Leeds, Leeds LS2 9JT, UK*

[*] Equal contribution

[1]Corresponding author. E-mails of corresponding authors: cafontes@fmv.ulisboa.pt willats@plen.ku.dk harry.gilbert@ncl.ac.uk

ABSTRACT

The breakdown of plant cell wall (PCW) glycans is an important biological and industrial process. Non-catalytic carbohydrate binding modules (CBMs) fulfill a critical targeting function in PCW depolymerization. Defining the portfolio of CBMs, the *CBMome*, of a PCW degrading system is central to understanding the mechanisms by which microbes depolymerize their target substrates. *Ruminococcus flavefaciens*, a major PCW degrading bacterium, assembles its catalytic apparatus into a large multienzyme complex, the cellulosome. Significantly, bioinformatic analyses of the *R. flavefaciens* cellulosome failed to identify a CBM predicted to bind to crystalline cellulose, a key feature of the CBMome of other PCW degrading systems. Here high throughput screening of 177 protein modules of unknown function was used to determine the complete CBMome of *R. flavefaciens*. The data identified six novel CBM families that targeted β-glucans, β-mannans and the pectic polysaccharide homogalacturonan. The crystal structures of four CBMs in conjunction with site-directed mutagenesis provides insight into the mechanism of ligand recognition. In the CBMs that recognize β-glucans and β-mannans differences in the conformation of conserved aromatic residues had a significant impact on the topology of the ligand binding cleft and thus ligand specificity. A cluster of basic residues in CBM77 confers calcium independent recognition of homogalacturonan indicating that the carboxylates of galacturonic acid are key specificity determinants. This report shows that the extended repertoire of proteins in the cellulosome of *R. flavefaciens* contributes to an extended CBMome that supports efficient PCW degradation in the absence of CBMs that specifically target crystalline cellulose.

Key words: Carbohydrate-Binding Modules, protein-carbohydrate interactions, Carbohydrate Active enZYmes, cellulosomes.

Significance Statement

Plant cell wall (PCW) polysaccharide degradation is an important biological and industrial process. Non-catalytic carbohydrate binding modules (CBMs) fulfill a critical targeting function in PCW depolymerization. *Ruminococcus flavefaciens* synthesizes a highly efficient PCW degrading apparatus. Here, six novel *R. flavefaciens* CBM families were identified that targeted β-glucans, β-mannans and pectins. Crystal structures of these CBMs revealed that recognition of β-glucans and β-mannans was mediated by differences in the conformation of conserved aromatic residues in the ligand binding cleft. A cluster of basic residues in CBM77 confers calcium independent recognition of homogalacturonan. This report shows that the expansion of proteins modules in the cellulosome of *R. flavefaciens* contributes to an extended CBM profile that supports efficient PCW degradation.

\body **INTRODUCTION**

Plant cell walls (PCWs) consist of interlinked polysaccharides, often impregnated with lignin that evolved to restrict access to enzyme attack. Thus, the recycling of photosynthetically fixed carbon is a slow biological process. Reflecting the intricacy of PCWs, microorganisms that degrade these composite structures produce extensive repertoires of carbohydrate active enzymes (CAZymes) (1), which are of increasing industrial significance (2).

CAZymes acting on recalcitrant carbohydrates are frequently appended with non-catalytic carbohydrate binding modules (CBMs). CBMs potentiate the activity of the associated catalytic modules through substrate targeting (see (3) for review). CBMs and CAZymes are classified into sequence-based families in the CAZy database (http://www.cazy.org/) (4). Based on their binding mode, CBMs have been classified into three types. Type A CBMs display a planar surface that binds to crystalline polysaccharides, Type B modules accommodate internal regions of glycan chains within open clefts, and Type C CBMs recognize the termini of glycans (exo-type) in a binding site that adopts a pocket topology (3).

Efficient hydrolysis of PCW polysaccharides has been fine-tuned over millions of years in ecological niches that are subjected to intensive selective pressures exemplified by the rumen of mammalian herbivores. A cohort of rumen anaerobic bacteria assemble their PCW degrading apparatus into multi-protein complexes termed cellulosomes (5). Cellulosome assembly is through the interaction of cohesin modules located on the non-catalytic protein, the scaffoldin, and dockerin modules on each enzyme subunit (5). Clostridial cellulosomes bind tightly to PCWs through a scaffoldin family 3 CBM. The repertoire of cellulosomal enzymes expressed by an individual bacterium constitutes a highly selected consortium of biocatalysts optimized to degrade PCWs. Genome sequencing of *Ruminococcus flavefaciens* strain FD-1 (6), the most abundant ruminal cellulolytic bacterium, revealed an elaborate assembly of scaffoldins indicating that the bacterium's cellulosome is an intricate and versatile PCW degrading system. Commensurate with this proposed cellulosomal complexity, the genome of *R. flavefaciens* FD-1 encodes ~230 dockerin containing proteins, which are likely to integrate into the multienzyme complex (6). A large number of the protein modules identified in the *R. flavefaciens* cellulosome are

of unknown function and may reflect an extended capacity to recognize carbohydrates through an extended CBM profile.

One of the major challenges facing postgenomic analysis of organisms is the identification of the function of the large number of predicted proteins derived from genomic sequencing. To bridge this gap in knowledge requires the development of high throughput methodologies (HTPMs). Here we have explored how HTPMs can be used to interrogate the functional complexity of the *R. flavefaciens* cellulosome. The data support the hypothesis that protein diversity in the *R. flavefaciens* cellulosome contributes to an expansion in glycan recognition, which is mediated by a ruminococcal-specific cohort of protein modules.

RESULTS AND DISCUSSION

R. flavefaciens cellulosomal enzymes contain novel CBMs

The *R. flavefaciens* FD-1 cellulosome contains 177 proteins modules of unknown function (UNKs). These UNKs were assessed for CBM functions using a carbohydrate microarray platform that enabled rapid screening of binding against multiple glycans. The microarrays were populated with 18 oligosaccharides and 46 polysaccharides of PCW origin (7). The output from the microarrays identified nine CBMs that bound to at least one arrayed glycan, **Fig. 1A** (non-binding glycans listed in **Table S1**). The positive hits were also screened by affinity gel electrophoresis (AGE), **Fig. 1B** and **Fig. S1A**, and the assignment of ligand specificity described below and in **Fig. 2** are derived from these AGE experiments. Based on sequence similarity the nine CBMs were grouped into six novel families designated CBM75 to CBM80. The protein modules are defined by the CBM family (CBMXX) and enzyme (RfGHXX) from which they are derived. Although the C-terminal portion of $CBM75_{RfGH43}$ displayed distant sequence similarity to members of CBM6, this region of the protein was not responsible for ligand recognition (see ***SI Results***), and thus the protein module was designated as a novel CBM family. An overview of the specificity of the six novel CBM families is as follows: CBM75 is a xyloglucan specific family, CBM76 recognized different β-1,4-glucans, CBM79 binds to a range of β-1,4- and mixed linked β-1,3-1,4-glucans. Similarly, families CBM78 and CBM80 display specificity for β-1,4- and mixed linked β-1,3-1,4-glucans, with some members also binding to β-1,4-mannans. Thus, in CBM78 and CBM80 the proteins $CBM78_{GH26}$ and

$CBM80_{RfGH5-1/2}$, respectively, bound to galactomannan in addition to the β-glucans, while $CBM78_{RfGH5}$ and $CBM80_{RfGH5}$ only recognized the *gluco*-configured ligands. None of the β-glucan binding CBMs bound to β-1,3-glucans, β-1,6-glucans or xylans. $CBM77_{PL1/9}$ bound exclusively to homopolygalacturonan (pectin) with low degrees of methyl esterification (DEs) *in vitro*, and to pectin within the context of intact PCWs, **Fig. 1C**.

The glycan microarray and AGE data were broadly similar, although subtle differences in specificity were evident. Thus, the binding of $CBM78_{RfGH5}$ to barley β-glucan was only observed using AGE, while only microarray data revealed an interaction between $CBM75_{RfGH43}$ and this glycan. Such differences in specificity between the two methods may reflect variations in the conformation of some glycans arrayed on nitrocellulose or contained within polyacrylamide gels.

The enzyme context of the novel CBMs

The CBMs that bound β-glucans are components of enzymes that contain catalytic modules derived from GH5_4 ($CBM78_{RfGH5}$, $CBM80_{RfGH5-1/2}$ and $CBM80_{RfGH5}$), GH9 (enzyme contains two CBM79s) or GH44 ($CBM76_{RfGH44}$), families/sub-families that are populated exclusively by endo-β1,4-glucanases, **Fig. 2**. The two enzymes containing CBMs that bind galactomannan have GH5_7 or GH26 "β1,4-mannanase" catalytic modules. Indeed, the dual specificity of $CBM80_{RfGH5-1/2}$ is consistent with the catalytic modules of the enzyme that hydrolyse β-glucans (GH5_4) or β-mannans (GH5_7). Consistent with its specificity, $CBM77_{RfPL1/9}$ is a component of an enzyme that contains two catalytic modules, belonging to polysaccharide lyase families 1 and 9 (PL1 and PL9), which display pectate lyase activity (k_{cat}/K_m values of PL1 and PL9 against homogalacturonan were 3.1×10^3 and 3.7×10^5 min^{-1} mg^{-1} ml, respectively). $CBM75_{RfGH43}$, which binds xyloglucan, is associated with a GH43_16 catalytic module, a subfamily that, to date, contains only arabinofuranosidases (8). The GH43 catalytic module of the enzyme hydrolyzed only 4-nitrophenyl-α-L-arabinofuranose demonstrating that the enzyme is an arabinofuranosidase. The enzyme was not active against arabinoxylans, and arabinans. CBMs generally display specificities consistent with the activity of the appended enzyme (3), although glycan recognition can be at the interface between the two modules illustrated by the arabinoxylan binding function of a modular arabinofuranosidase (9). Given that $CBM75_{RfGH43}$ binds to xyloglucan, we speculate that the GH43_16 targets arabinofuranose residues that decorate xyloglucans from tomato (10).

Phylogeny of the six novel CBM families

Representatives of the six novel *Ruminococcus* CBM families were used in a BLAST search of the NCBI protein sequence database. Sequences were retrieved with *E* values $<4 \times 10^{-4}$ with sequence identify >30%. No sequence corresponding to a CBM annotated on the CAZy database (4) was identified, confirming the discovery of six novel CBM families, **Fig. S2**. Families CBM75-76 and CBM78-80 contain sequences derived exclusively from ruminococci. In contrast, the catalytic modules appended to these CBMs display sequence similarity to glycoside hydrolases from a range of prokaryotes. These data indicate that the catalytic modules of the *R. flavefaciens* enzymes were acquired through horizontal gene transfer and subsequently appended with the *Ruminococcus* specific CBMs. CBM77 contains ~140 members from a range of bacteria. In the majority of the non-ruminococci proteins the CBM77 is appended to PL1 catalytic modules, although PL9 sequences were also present in a cohort of enzymes, **Fig. S3**. Based on the phylogeny, it would appear that CBM75, 76, 78, 79, and 80 fulfill an enzyme targeting role that is specific to *Ruminococcus*. It is possible that the contribution of these CBMs to enzyme function is only evident in a highly complex scaffold provided by the intricate organization of the *R. flavefaciens* cellulosome. In contrast, the CBM77 appears to play a more general role in pectin degradation that is not specific to ruminococci or cellulosome organization.

Thermodynamics of ligand binding for selected CBMs

The affinity of representatives of the six novel CBM families for their respective ligands was determined by isothermal titration calorimetry (ITC), **Table 1** and **Fig. S1B,** with thermodynamic parameters reported in **Table S2**. $CBM75_{GH43}$ bound exclusively to xyloglucan. The affinity of the CBM for the oligosaccharide XXXG (X, glucose linked O6 to xylose; G, undecorated glucose), the repeating unit of xyloglucan, was similar to the polysaccharide, **Table 1**. This indicates that the protein contains four binding sites that interact with the backbone glucose units and at least some of the xylose side chains.

$CBM76_{GH44}$ and $CBM78_{GH5}$ displayed highest affinity for xyloglucan. The similar K_A values of $CBM78_{GH5}$ for cellohexaose and cellopentaose suggested five dominant sugar binding sites. The affinity of $CBM78_{GH5}$ for XXXG was significantly higher than

cellotetraose, **Table 1**, suggesting that recognition of the xylose side chains occurred within the core binding sites of the protein.

CBM77$_{PL1/9}$ displayed high affinity for low DE pectins, **Table 1**. Only oligosaccharides with a degree of polymerization (DP) ≥7 bound to the CBM, indicating that the binding site interacts with 7 or 8 GalA residues. Ligand recognition by CBM77$_{PL1/9}$ was not inhibited by EDTA, indicating that pectin binding was metal independent. This is in contrast to pectate lyases where calcium is a central feature of GalA recognition. CBM32 from *Yersinia* (11), which is not a component of an enzyme, is the only other example of a CBM that binds to pectin backbones. The CBM32 binds optimally to ~10 GalA residues, although the role of metal ions in ligand recognition was not reported.

CBM79-1$_{GH9}$ (N-terminal CBM79) bound to barley β-glucan and HEC with similar affinities. The small increase in K_A from cellotetraose to cellohexaose, **Table 1**, suggests that ligand recognition is again dominated by four tandem sugar binding sites. Binding to xyloglucan was weaker than the other β-glucans, indicating that the protein cannot easily accommodate the xylose side chains of the hemicellulose.

CBM80$_{GH5-1/2}$ bound to β-glucans and galactomannan with affinities in the range of 10^4 to 10^5 M^{-1}, **Table 1**. The protein bound to cellulooligosaccharides and mannooligosaccharides with a DP of 5 or 6 with similar affinities. While CBM80$_{GH5-1/2}$ bound to mannotetraose, the protein did not bind to cellotetraose. Thus, the binding region of CBM80$_{GH5-1/2}$ for the *gluco*-configured ligands is more extensive than for the mannose-based glycans (see below).

Crystal Structures and ligand recognition of representatives of CBM77$_{RfPL1/9}$, CBM78$_{RfGH5}$, CBM79-1$_{RfGH9}$ and CBM80$_{RfGH5-1/2}$

The crystal structures of representative CBMs that target pectin, β-glucans and mannan/β-glucan, respectively, were solved, to a resolution of 1.5 Å (CBM77$_{RfPL1/9}$), 2.0 Å (CBM78$_{RfGH5}$), 1.8 Å (CBM79-1$_{RfGH9}$) and 1.0 to 1.5 Å (CBM80$_{RfGH5-1/2}$); structure statistics in **Table S3**. All the proteins adopt a β-sandwich fold, **Fig. 3**, typical of CBM families (3). The *R. flavefaciens* proteins displayed 3D structural similarity, but very low sequence identity (3-11%), to proteins in other CAZy CBM families, **Table S4**. There is, however, no conservation in the ligand binding residues between the *Ruminococcus* proteins and the structural homologs in the other CAZy

CBM families. Details of the secondary structures of the CBMs are provided in **Fig. S4** and the structural basis for ligand recognition are as follows:

CBM77$_{RfPL1/9}$: CBM77$_{RfPL1/9}$ contains two β-sheets defined as 1 and 2, **Fig. 3D and S4**: The canonical ligand binding site in endo-acting type B CBMs comprise the concave surface presented by β-sheet 2 (3). In CBM77$_{RfPL1/9}$, however, β-sheet 2 does not display this classic cleft topology, and mutation of aromatic and basic residues in the concave surface had no effect on ligand binding, **Table S2D**. The surface of CBM77$_{RfPL1/9}$, **Table S4,** comprising the loops connecting the β-sheets contains indentations. At the centre of this surface are Lys^{1092}, Lys^{1107} and Lys^{1162}. Alanine substitution of these residues abrogated pectin showing that these amino acids comprise the ligand binding site, **Table S2D**. Distal to the central basic ligand binding site are two additional lysine residues, Lys^{1136} and Lys^{1141}. The double mutant K1136A/K1141A displayed no binding to pectin, although the individual mutants, K1136A and K1141A, retained affinity for pectin, **Table S2D**. Lys^{1136} and Lys^{1141} may only bind pectin when the polysaccharide is in one of its two possible orientations, explaining the functional redundancy. The ligand binding surface is ~25 Å. Pectic homogalacturonan adopts a compressed "accordion-like" structure in which a disaccharide subunit spans a distance of 8 Å (11), suggesting that the binding site can accommodate a hexasaccharide. The ligand binding mode of CBM77$_{RfPL1/9}$ is distinct from other CBMs where aromatic residues dominate glycan recognition (3), but resembles glycosaminoglycan binding proteins where ligand recognition is also mediated by basic residues (12). Within the CBM77 family the three core pectin binding residues in CBM77$_{RfPL1/9}$ (Lys^{1092}, Lys^{1107} or Lys^{1162}) were invariant, **Fig. S2**. Thus, pectin recognition appears to be a conserved feature of CBM77.

CBM78$_{RfGH5}$ and* CBM79-1$_{RfGH9}$ *representatives of CBMs that bind β-glucans: CBM78$_{RfGH5}$ and CBM79-1$_{RfGH9}$ contain two β-sheets defined as 1 and 2, respectively, **Fig. S4**. In both CBMs β-sheet 2 adopts forms a cleft in which aromatic residues are a dominant feature, **Fig 3**. In CBM78$_{RfGH5}$ Trp^{496}, Trp^{554}, Tyr^{555} and Phe^{479} are aligned along the cleft, while in CBM79-1$_{RfGH9}$ Tyr^{563}, Trp^{564}, Tyr^{597}, Trp^{606} and Trp^{607} form a twisted hydrophobic platform. These hydrophobic regions are predicted to comprise the glucan binding sites in the two CBMs, **Fig. 3A,B**.

Mutagenesis confirmed the importance of the aromatic residues in ligand recognition in β-sheet 2 of the two proteins. Alanine substitution of Trp^{496} or Trp^{554} in $CBM78_{RfGH5}$, and Trp^{606} in $CBM79\text{-}1_{RfGH9}$, which are conserved in the two CBM families, **Fig. S2**, resulted in complete loss of binding to all ligands, **Table S2A,B**. A significant feature of these mutagenesis experiments was that alanine substitutions of several residues modulated ligand specificity. With respect to $CBM78_{RfGH5}$, the mutants F479A and Y555A bound to xyloglucan but not to barley β-glucan or HEC. The variant Q552A recognized xyloglucan and barley β-glucan but not HEC, **Table S2A**. The equivalent residue to $CBM78_{RfGH5}$ Phe^{479} is not aromatic in four members of family CBM78, **Fig. S2**, suggesting that these proteins may bind to xyloglucan but not to other β-glucans. In $CBM79\text{-}1_{RfGH9}$ the mutants W564A and W607A retained affinity for barley β-glucan but did not bind xyloglucan, **Table S2B**. Trp^{607} in $CBM79\text{-}1_{RfGH9}$ is replaced by glycine in two members of CBM79, **Fig. S2**, suggesting that there are also differences in specificity within this family. To summarize, the mutagenesis data show that while core residues play a generic role in binding β-glucans, other amino acids in the two CBMs play distinct roles in ligand recognition, explaining why these proteins can bind to a range of β-glucans.

The topology of the ligand binding site of $CBM78_{RfGH5}$ and $CBM79\text{-}1_{RfGH9}$ are very different, even though the positions of the key glucan binding aromatic residues are conserved. In $CBM78_{RfGH5}$ the cleft is a narrow canyon-like structure. In $CBM79\text{-}1_{RfGH9}$, however, the concave surface forms an unusually solvent exposed cleft or planar surface with loops connecting β-strands 1 and 2 and β-strands and 4 and 5 strongly contributing to the curved topology of β-sheet 2, **Fig. 3A,B**. These contrasting topologies reflect the orientation of two tryptophan residues that play a key role in ligand recognition. With respect to β-sheet 2 these residues adopt a planar (Trp^{564} and Trp^{606}) or perpendicular (Trp^{496} and Trp^{554}) orientation in $CBM79\text{-}1_{RfGH9}$ and $CBM78_{RfGH5}$, respectively, **Fig. 3E**. In $CBM78_{RfGH5}$ β-sheet 2 contains two additional β-strands (7 and 8), and Ile^{522} from strand 7 stacks against the indole ring of Trp^{496}, enabling the aromatic residue to adopt a perpendicular orientation. The Nε of Trp^{554} makes a polar contact with Oε1 of Gln^{552}, which orientates the indole ring into its perpendicular conformation. In contrast, there are no steric constraints preventing Trp^{564} and Trp^{606} in $CBM79\text{-}1_{RfGH9}$ from making apolar planar interactions with the peptide chain of β-strands 4/7 and 4/5, respectively.

The planar topology of the binding cleft of CBM79-1$_{RfGH9}$ indicates that this protein may interact with components of insoluble cellulose. The narrow binding cleft of CBM78$_{RfGH5}$ points to a specificity for discrete cellulose chains, which occur rarely in insoluble cellulose. ITC and pull down assays showed that CBM79-1$_{RfGH9}$, but not CBM78$_{RfGH5}$, bound to regenerated (non-crystalline) insoluble cellulose (RC), **Fig. S1C,D**. The inability of W564A and W606A to bind RC shows that recognition of the polysaccharide is in solvent exposed cleft, **Fig. S1D**. The impact of the conformation adopted by conserved aromatic residues on CBM specificity, is also apparent in family 2 CBMs that bind to cellulose or xylan (13). The differences in the specificity of β-glucan binding CBMs reported here provide a biological rationale for the evolution of a large number of CBMs that target these glycans. Within the context of a complex cellulosome structure the myriad of β-glucan binding CBMs may act in synergy to bind different substructures of cellulose, obviating the need for a classical type A module that binds crystalline cellulose.

CBM80$_{RfGH5-1/2}$: The apo structure of CBM80$_{RfGH5-1/2}$ and in complex with mannohexaose and cellohexaose were solved to a resolution of 1.0 Å, 1.4 Å and 1.5 Å, respectively. The β-sheet 2 of CBM80$_{RfGH5-1/2}$ presents a planar hydrophobic surface, through the approximately parallel orientation of Trp^{453} and Trp^{489}. Distal to this planar surface is a third aromatic residue, Trp^{490}, which is in a perpendicular orientation to the two other solvent exposed tryptophan residues.

The mannohexaose-CBM80$_{RfGH5-1/2}$ complex revealed electron density for mannohexaose, **Fig. S5A**, along the hydrophobic surface of β-sheet 2. All of the pyranoside sugars were in the undistorted $^4C_1$ chair conformation. The three solvent exposed aromatic residues (Trp^{453}, Trp^{489} and Trp^{490}) interact with the α-face of the pyranose rings of the mannoside residues 2, 4, and 6, respectively. The minimum ligand required to harness the binding energy from all three aromatics is a pentasaccharide, consistent with the ITC binding data, **Table 1.** There are few direct hydrogen bonds between CBM80$_{RfGH5-1/2}$ and mannohexaose, **Fig. 3**. Mannoside-1 (reducing end sugar) and mannoside-4 make no direct hydrogen bonds to the CBM. The O3 of mannoside-2 forms hydrogen bonds with Oε1 of Glu^{485} and Nε2 of Gln^{487} makes a polar contact with O3 and O4. Mannoside-3 interacts with the protein through polar contacts between O2 and both Oε1 of Gln^{487} and Nε1 of Trp^{489}, O3 and Nζ1 of Lys^{455}, while the endocyclic oxygen accepts a hydrogen from Nε2 of

Gln^{487}. O2 of mannoside-5 makes a hydrogen bond with Oδ1 of Asn^{442}, while Nδ2 of the asparagine makes polar contacts with O3 of mannosides 5 and 6, respectively. The O6 is solvent exposed at mannosides-1 to 5, and thus CBM80$_{RfGH5-1/2}$ is able to recognize the backbone of galactomannans in which the mannan backbone is decorated with α-1,6-D-galactose side chains. The structure of CBM80$_{RfGH5-1/2}$ in complex with cellohexaose revealed electron density for only three glucose units, which were also modelled in their relaxed $^4C_1$ conformation, **Fig. S5B**. The cellulooligosaccaride and the corresponding region of mannohexaose made similar interactions with the CBM, **Fig. S5C**. The only differences were that O2 of glucoside-2 and glucoside-3 interacted with Oε1 of Glu^{485} and Nζ1 of Lys^{455}, respectively, while Gln^{487} makes three less polar contacts with the *gluco*-configured ligand compared to the mannooligosaccharide.

AGE showed that the mutants K455A, E485A and Q487A of CBM80$_{RfGH5}$ retained wild type affinity for all the ligands tested. This suggests that the predicted polar interactions between the protein and β-glycans have very little influence on affinity. These data are unusual among type B CBMs where direct polar interactions generally make an important contribution to ligand recognition (14, 15). The data, however, are consistent with *Cf*CBM2b-1, in which affinity is dominated by two tryptophans (16). Although removal of polar interactions greatly reduced the enthalpy of binding, because of enthalpy-entropy compensation the mutations did not influence K_A values, which likely reflect the exposed binding site presented by this protein. CBM80$_{RfGH5-1/2}$ also contains a highly exposed ligand binding site and thus the retention of wild type affinity in the polar mutants may also result from enthalpy-entropy compensation. The distinguishing feature of glucose and mannose is the stereochemistry at O2, which adopts an equatorial or axial conformation, respectively. The observation that Q487A, K455A and E457A substitutions did not influence affinity for β-glucans or β-mannans indicates that O2 is not a significant specificity determinant for these ligands. This explains why CBM80$_{RfGH5}$ is able to bind to both cellulose and β-mannan. Examples of CBMs that recognise both β-1,4-glucans and β-1,4-mannans are found in families CBM16 (17) and CBM29 (CBM29-2) (18). In both proteins residues that interact with O2 can make hydrogen bonds with the hydroxyl in either its axial or equatorial conformation. As discussed above, this is in contrast to CBM80$_{RfGH5-1/2}$ where polar interactions with its ligands do not appear to contribute to affinity or specificity.

A key discriminator of ligand specificity is likely to be the location of the three aromatic residues in the binding cleft that exhibit specificity for polysaccharide chains displaying a "twisted" conformation. This conformation has been shown for cellohexaose in solution (19) and when bound to other CBMs (20). Given that $CBM80_{RfGH5\text{-}1/2}$ does not target the O6 groups in hexasaccharides it is perhaps surprising the CBM does not bind xylan. This likely reflects a binding cleft that is optimized to bind a twisted β-1,4-glycan chain, and is unable to accommodate the conformation adopted by xylan, a helical glycan with a three-fold screw axis.

The importance of the surface tryptophan residues in ligand recognition is revealed by the complete abrogation of binding to β-glucans and β-mannans when Trp^{453} and Trp^{489}, which stack against mannoside and glucoside residues 2 and 4, were substituted with alanine, **Table S2C**. The mutant W490A, retained the capacity to bind β-1,4-mannans, albeit with a 10-fold reduction in K_A, but displayed no affinity for the β-glucans, **Table S2C**. CBMs that bind to extensive β-1,4-glycans typically contain three aromatic residues that make apolar interactions with sugars *n, n+2* and *n+4* (18). Alanine substitution of any of these aromatic residues generally leads to a substantial reduction and often complete abrogation of binding (21). It is unique, therefore, to observe a differential effect on glycan recognition when one of these aromatic residues is substituted with alanine. It is not obvious why the W490A mutant introduced selectivity for *manno*-configured ligands. The crystal structure of the mutant revealed no significant conformational changes with an RMSD of 0.7 Å compared to the wild type CBM, **Fig. S5D**. A notable feature of Trp^{490} is that the side chain is orientated ~90^{o} compared to the other two surface tryptophan residues. We speculate that Trp^{490} contributes less binding energy for mannan compared to glucan, as the *manno*-configured ligand is required to adopt non-optimal conformations to access the distal aromatic residue. Thus, substitution of Trp^{490} has less impact on mannan binding than β-1,4-glucan recognition. This would suggest that although both mannooligosaccharides and cellulooligosaccharides display a twisted conformation when bound to a variety of CBMs, their minimum energy conformations are not identical.

Conclusion

In the last decade, the availability of genomic/metagenomic data has increased exponentially. It is apparent that HTPMs need to be developed to understand the

biological and biotechnological significance of this explosion in sequence information. This study explored the use of microarray technology combined with HTP protein production to explore the function of the 177 *R. flavefaciens* UNKs that comprise 50% of the subunits of the most complex cellulosome described to date. The data revealed six novel CBM families of which five target β-glucans and/or β-mannan, and one recognizes homogalacturonan. Structural data, in addition to revealing the importance of basic residues in calcium independent pectin recognition, showed how the conformation of conserved aromatic residues can have a profound influence on the topology of the substrate binding cleft and consequently influence specificity.

To conclude, the data reported here reveals an unprecedented expansion in glycan recognition by the cellulosomes of rumen bacteria specialized in PCW degradation. This indicates that in highly competitive ecological niches, where complex carbohydrates are used as a major carbon source, enzyme–substrate targeting through the function of CBMs plays a critical role in substrate acquisition and thus organism survival.

MATERIALS AND METHODS

Cloning, expression, site-directed mutagenesis and purification of cellulosomal UNKs

The genes encoding 177 cellulosomal UNKs from *R. flavefaciens* were cloned into *Escherichia coli* expression vectors. Details of the cloning strategies, site-directed mutagenesis and the purification of the recombinant proteins are described in ***SIMethods***,

Glycan binding assays

The binding of CBMs to carbohydrate ligands was screened using carbohydrate microarrays printed on nitrocellulose and AGE The binding of selected CBMs to their ligands was quantified by ITC and insoluble ligands by pull down experiments, and are described in detail in ***SI Methods***

Crystallization, Data Collection, and Structure Solution

The structures of $CBM78_{RfGH5}$, $CBM79\text{-}1_{RfGH9}$ and $CBM80_{RfGH5\text{-}1/2}$ were solved using SAD methods and selenomethionyl proteins, and $CBM77_{RfPL1/9}$ was solved by sulphur

SAD methods. Details of crystallization, data collection, and structure solution, are given in ***SI Methods***. Structure statistics are provided in **Table S3**.

Acknowledgments

This work was supported by the EU FP7 programme under the WallTraC project (grant No. 263916) and BioStruct-X (grant No. 283570), by Danish Strategic Research Council and The Danish Council for Independent Research, Technology and Production Sciences as part of the GlycAct project (FI 10-093465) and Qren through grant 30263. We thank Diamond Light Source (Harwell, UK) for access to beamlines I02, I04 and I04-1 (mx9948) that contributed to the results presented here as well as the European Synchrotron Radiation Facility (Grenoble, France) and Soleil (Saint-Aubin, France) for data collection.

REFERENCES

1. Gilbert HJ (2010) The biochemistry and structural biology of plant cell wall deconstruction. *Plant Physiol.* 153(2):444-455.
2. Demain AL, Newcomb M, Wu JH (2005) Cellulase, clostridia, and ethanol. *Microbiol Mol Biol Rev* 69(1):124-154.
3. Gilbert HJ, Knox JP, Boraston AB (2013) Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin Struct Biol* 23(5):669-677.
4. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42(Database issue):D490-495.
5. Bayer EA, Belaich JP, Shoham Y, Lamed R (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* 58:521-554.
6. Rincon MT*, et al.* (2010) Abundance and diversity of dockerin-containing proteins in the fiber-degrading rumen bacterium, Ruminococcus flavefaciens FD-1. *PloS one* 5(8):e12476.
7. Pedersen HL*, et al.* (2012) Versatile high resolution oligosaccharide microarrays for plant glycobiology and cell wall research. *J Biol Chem* 287(47):39429-39438.
8. Mewis K, Lenfant N, Lombard V, Henrissat B (2016) Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization. *Appl. Environ. Microbiol.* 82:*in the press*.
9. Yoshida S, Hespen CW, Beverly RL, Mackie RI, Cann IK (2010) Domain analysis of a modular alpha-L-Arabinofuranosidase with a unique carbohydrate binding strategy from the fiber-degrading bacterium Fibrobacter succinogenes S85. *J Bacteriol* 192(20):5424-5436.
10. Jia Z, Qin Q, Darvill AG, York WS (2003) Structure of the xyloglucan produced by suspension-cultured tomato cells. *Carbohydrate research* 338(11):1197-1208.
11. Abbott DW, Hrynuik S, Boraston AB (2007) Identification and characterization of a novel periplasmic polygalacturonic acid binding protein from *Yersinia enterolitica*. *J Mol Biol* 367(4):1023-1033.

12. Hileman RE, Fromm JR, Weiler JM, Linhardt RJ (1998) Glycosaminoglycan-protein interactions: definition of consensus sites in glycosaminoglycan binding proteins. *Bioessays* 20(2):156-167.
13. Simpson PJ, Xie H, Bolam DN, Gilbert HJ, Williamson MP (2000) The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J Biol Chem* 275(52):41137-41142.
14. Pell G*, et al.* (2003) Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry* 42(31):9316-9323.
15. Su X*, et al.* (2010) Mutational insights into the roles of amino acid residues in ligand binding for two closely related family 16 carbohydrate binding modules. *J Biol Chem* 285(45):34665-34676.
16. Xie H*, et al.* (2001) Role of hydrogen bonding in the interaction between a xylan binding module and xylan. *Biochemistry* 40(19):5700-5707.
17. Bae B*, et al.* (2008) Molecular basis for the selectivity and specificity of ligand recognition by the family 16 carbohydrate-binding modules from *Thermoanaerobacterium polysaccharolyticum* ManA. *J Biol Chem* 283(18):12415-12425.
18. Charnock SJ*, et al.* (2002) Promiscuity in ligand-binding: The three-dimensional structure of a *Piromyces* carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. *Proc. Natl. Acad. Sci. U S A* 99(22):14077-14082.
19. Sugiyama H, Hisamichi K, Usui T, Sakai K, Ishiyama J (2000) A study of the conformation of beta-1,4-linked glucose oligomers, cellobiose to cellohexaose, in solution. *J Mol Struct* 556(3):173-177.
20. Boraston AB*, et al.* (2002) Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. *J Mol Biol* 319(5):1143-1156.
21. Nagy T*, et al.* (1998) All three surface tryptophans in Type IIa cellulose binding domains play a pivotal role in binding both soluble and insoluble ligands. *FEBS Lett* 429(3):312-316.
22. Boraston AB (2005) The interaction of carbohydrate-binding modules with insoluble non-crystalline cellulose is enthalpically driven. *Biochem J* 385(Pt 2):479-484.
23. Charnock SJ*, et al.* (2000) The X6 "thermostabilizing" domains of xylanases are carbohydrate-binding modules: structure and biochemistry of the *Clostridium thermocellum* X6b domain. *Biochemistry* 39(17):5013-5021.
24. Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):125-132.
25. Evans PR,Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* 69(Pt 7):1204-1214.
26. Terwilliger TC*, et al.* (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr D Biol Crystallogr* 65(Pt 6):582-601.
27. McCoy AJ*, et al.* (2007) Phaser crystallographic software. *J Appl Crystallogr* 40(Pt 4):658-674.
28. Sheldrick GM (2010) Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr D Biol Crystallogr* 66(Pt 4):479-485.
29. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66(Pt 4):486-501.
30. Murshudov GN*, et al.* (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 67(Pt 4):355-367.
31. Winn MD*, et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67(Pt 4):235-242.

32. Chen VB, *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66:12-21.
33. Correia MA, *et al.* (2011) Structure and function of an arabinoxylan-specific xylanase. *J Biol Chem* 286(25):22510-22520.

FIGURE LEGENDS

Fig 1. Screening *R. flavefaciens* UNKs for glycan binding functions. ***Panel A:*** Glycan microarray binding profiles of the founding members of six novel CBM families. CBM6, 30 and 3a are characterized CBMs and JIM15, a monoclonal antibody (mAb) that binds homogalacturonan, were used as binding controls.The mean spot signals obtained from two individual experiments is presented in a heat map in which color intensity is correlated to signal. The highest signal in the data set was set to 100 and all other values were normalized accordingly (in accordance with the color intensity scale bar). Glycans that did not bind to any of the proteins screened are listed in **Table S1**. ***Panel B:*** Binding affinity of different CBMs detected by AGE. Red signifies binding, light green marginal binding and blue represents no detectable binding. ***Panel C:*** Evaluation of the binding of $CBM77_{RfPL1/9}$ to pectic homogalacturonan in tobacco stem sections. LM19 is a mAb that binds unesterified homogalacturonan. The binding capacity of $CBM77_{RfPL1/9}$ and LM19 were evaluated before and after section pre-treatment with a pectate lyase.

Fig. 2. Molecular architectures of proteins containing novel CBMs identified in this study. The origin of the families of glycoside hydrolases (GH) and polysaccharide lyases (PL) are identified. The modules of unknown function are colored grey, signal peptides red and dockerin modules purple. Linker sequences are depicted by a line. The boundaries of the modules in the full-length sequence of the enzymes are indicated. The polysaccharides targeted by the different CBMs are indicated.

Fig. 3. Crystal structures of $CBM79\text{-}1_{RfGH9}$, $CBM78_{RfGH5}$, $CBM80_{RfGH5\text{-}1/2}$ and $CBM77_{RfPL1/9}$. *Panels A, B, C* and *D* depict schematics of $CBM78_{RfGH5}$, $CBM79\text{-}1_{RfGH9}$, $CBM80_{RfGH5\text{-}1/2}$ and $CBM77_{RfPL1/9}$, respectively, color ramped from N terminus (blue) to C terminus (red), embedded in the surface representation of the proteins. The aromatic ($CBM78_{RfGH5}$, $CBM79\text{-}1_{RfGH9}$, $CBM80_{RfGH5\text{-}1/2}$) or basic ($CBM77_{RfPL1/9}$) residues that contribute to ligand recognition are shown in stick format and identified in *Panels A, B* and *D*. *Panel E* displays an overlay of $CBM78_{RfGH5}$ (green) and $CBM79\text{-}1_{RfGH9}$ (magenta). The panel also shows the residues that interact with the aromatic amino acids in $CBM78_{RfGH5}$. *Panel F* and *G* shows the amino acids (carbons

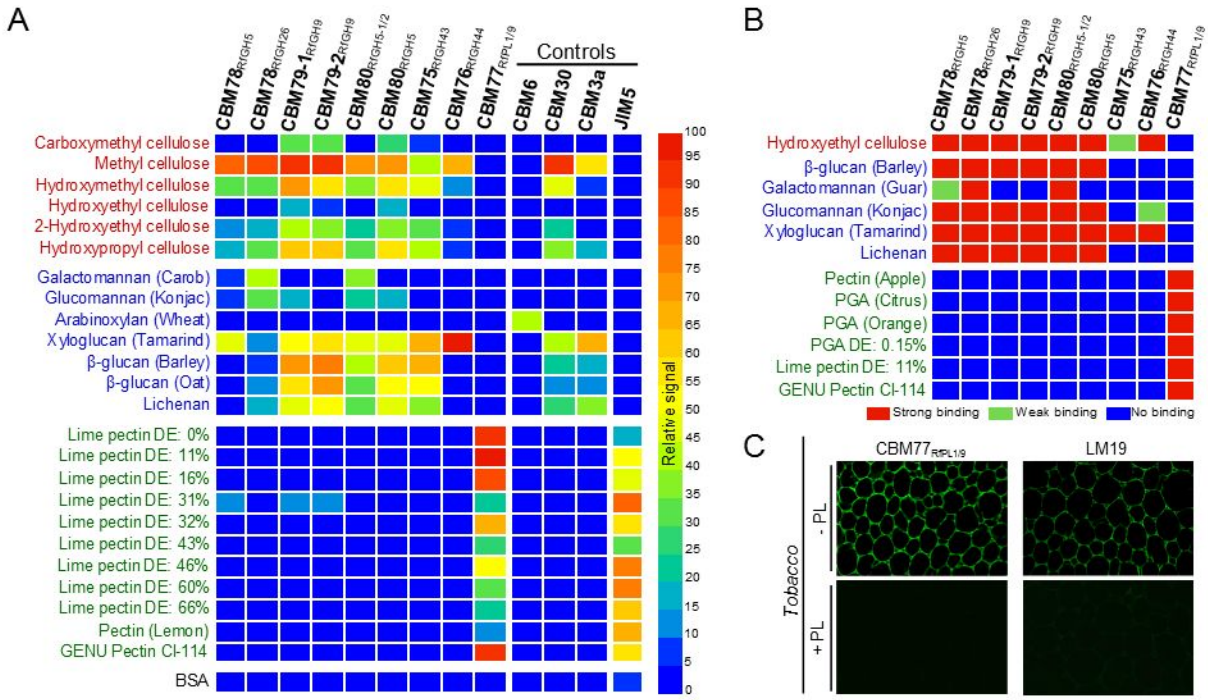
colored *green*) in the structure of $CBM80_{RfGH5\text{-}1/2}$ that make polar (indicated by *black dashed lines*) or apolar interactions with mannohexaose (carbons colored *yellow*). The ligands are labelled from the reducing end (i.e. Man-1). This figure and Fig. S5 were prepared using PyMol.

Table 1. The affinity of CBMs for their ligands

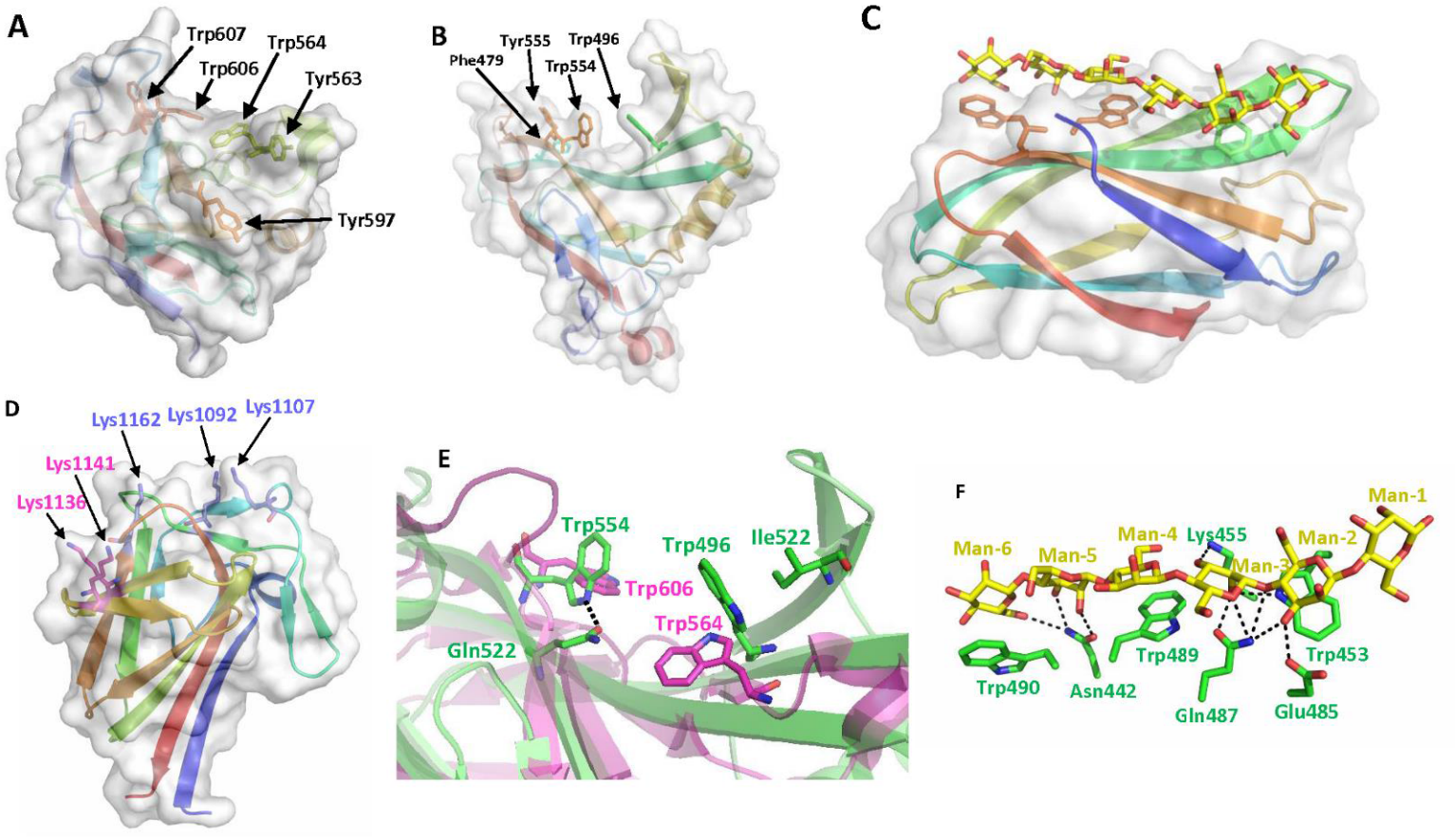
CBM	Ligand	K_A, M^{-1}
$CBM75_{RfGH43}$	Xyloglucan	1.7 (±0.2) x 10^4
	Glucomannan	No binding
	β-Glucan	No binding
	HEC[a]	Binding weak[d]
	XXXG[b]	4.0 (±0.6) x 10^3
$CBM76_{RfGH44}$	Xyloglucan	1.1 (±0.0) x 10^6
	Glucomannan	3.8 (±0.2) x 10^4
	β-Glucan	1.2 (±0.1) x 10^4
	HEC[a]	2.6 (±0.1) x 10^4
	XXXG[b]	1.6 (±0.2) x 10^4
$CBM77_{RfPL1/9}$	Lime Pectin DE 11%	1.8 (±0.2) x 10^5
	GENU pectin CI-114	4.2 (±0.2) x 10^5
	PGA from orange	1.2 (±0.02) x 10^4
	PGA from orange + EDTA	1.1 (±0.0) x 10^4
	Pectin from citrus DE 30%	1.1 (±0.0) x 10^4
	Pectin from citrus DE 60%	Binding weak[d]
	Pectin from citrus DE ≥80%	No binding
	GalA DP3/DP4	Binding weak[d]
	GalA DP7/DP8	1.2 (±0.2) x 10^5
$CBM78_{RfGH5}$	Xyloglucan	1.4 (±0.1) x 10^5
	β-Glucan	2.4 (±0.4) x 10^3
	HEC[a]	2.1 (±0.2) x 10^4
	XXXG[b]	3.0 (±0.7) x 10^3
	Cellohexaose	1.7 (±0.1) x 10^4
	Cellopentaose	8.6 (±0.2) x 10^3
	Cellotetraose	Binding weak[d]
	RC[c]	No binding
$CBM79\text{-}1_{RfGH9}$	Xyloglucan	1.0 (±0.2) x 10^4
	β-Glucan	4.0 (±0.2) x 10^4
	HEC[a]	7.8 (±0.5) x 10^4
	XXXG[b]	No binding
	Cellohexaose	4.9 (±0.9) x 10^3
	Cellopentaose	7.0 (±0.3) x 10^3
	Cellotetraose	4.2 (±1.7) x 10^3
	RC[c]	4.8 (±0.2) x 10^4
$CBM80_{RfGH5\text{-}1/2}$	Xyloglucan	1.8 (±0.1) x 10^5
	Glucomannan	5.8 (±0.5) x 10^4
	Galactomannan	4.5 (±0.2) x 10^4
	β-Glucan	2.2 (±0.2) x 10^4
	HEC[a]	3.6 (±0.1) x 10^3
	XXXG[b]	No binding
	Mannohexaose	4.1 (±0.9) x 10^4
	Mannopentaose	2.9 (±0.4) x 10^4
	Mannotetraose	1.8 (±0.2) x 10^3
	Cellohexaose	1.7 (±0.2) x 10^4
	Cellopentaose	8.5 (±0.7) x 10^3
	Cellotetraose	No binding

Affinities were determined by isothermal titration calorimetry. The thermodynamics of ligand binding are reported in **Table S2**.

[a]HEC, hydroxyethyl cellulose; [b]XXXG, xyloglucan heptasaccharide; [c]RC, regenerated cellulose; [d]binding too weak to quantify by ITC.


A
CBM78RtGH5
CBM78RtGH26
CBM79-1RtGH9
CBM79-2RtGH9
CBM80RtGH5-1/2
CBM80RtGH5
CBM75RtGH43
CBM76RtGH44
CBM77RtPL1/9
Controls
CBM6
CBM30
CBM3a
JIM5
Carboxymethyl cellulose
Methyl cellulose
Hydroxymethyl cellulose
Hydroxyethyl cellulose
2-Hydroxyethyl cellulose
Hydroxypropyl cellulose
Galactomannan (Carob)
Glucomannan (Konjac)
Arabinoxylan (Wheat)
Xyloglucan (Tamarind)
β-glucan (Barley)
β-glucan (Oat)
Lichenan
Lime pectin DE: 0%
Lime pectin DE: 11%
Lime pectin DE: 16%
Lime pectin DE: 31%
Lime pectin DE: 32%
Lime pectin DE: 43%
Lime pectin DE: 46%
Lime pectin DE: 60%
Lime pectin DE: 66%
Pectin (Lemon)
GENU Pectin CI-114
BSA
Relative signal
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
B
CBM78RtGH5
CBM78RtGH26
CBM79-1RtGH9
CBM79-2RtGH9
CBM80RtGH5-1/2
CBM80RtGH5
CBM75RtGH43
CBM76RtGH44
CBM77RtPL1/9
Hydroxyethyl cellulose
β-glucan (Barley)
Galactomannan (Guar)
Glucomannan (Konjac)
Xyloglucan (Tamarind)
Lichenan
Pectin (Apple)
PGA (Citrus)
PGA (Orange)
PGA DE: 0.15%
Lime pectin DE: 11%
GENU Pectin CI-114
Strong binding
Weak binding
No binding
C
CBM77RtPL1/9
LM19
Tobacco
- PL
+ PL

ENZYME	CBM	MOLECULAR ARCHITECTURE	PRIMARY LIGANDS
RfGH43	$CBM75_{RfGH43}$		Xyloglucan
RfGH44	$CBM76_{RfGH44}$		β-1,4-glucans Xyloglucan Glucomannan
RfPL1/9	$CBM77_{RfPL1/9}$		Pectins
RfGH5	$CBM78_{RfGH5}$		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan
RfGH26	$CBM78_{RfGH26}$		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan β-1,4-mannans
RfGH9	$CBM79_{RfGH9}$		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan
RfGH5-1/2	$CBM80_{RfGH5-1/2}$		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan β-1,4-mannans
RfGH5	$CBM80_{RfGH5}$		β-1,4-glucans β-1,3-1,4-glucans Xyloglucan Glucomannan


A
Trp607
Trp606
Trp564
Tyr563
Tyr597
B
Phe479
Tyr555
Trp554
Trp496
C
D
Lys1162
Lys1092
Lys1107
Lys1141
Lys1136
E
Trp554
Trp496
Ile522
Trp606
Trp564
Gln522
F
Man-1
Man-6
Man-5
Man-4
Lys455
Man-2
Man-3
Trp489
Trp453
Trp490
Asn442
Gln487
Glu485